Data Science Update:
Male Family Planning Analysis

•••

Sophia Atik
Applied Data Science
April 24th, 2019

Chosen Dataset

2015-2017 National Survey of Family Growth
MALE Questionnaire

Chosen Dataset

- Males age 15 to 49
- 4,540 Males Interviewed
- 2,945 variables
- Interviewed between Sept. 2015- Sept. 2017
- 60 min interviews under Research Ethics Review Board Guidelines (protocol #2015-12)
- Interviewees were incentivized with $40

MALE RESPONDENT FILE *– Information for each male respondent*

- Respondent ID (CASEID) and selected screener variables
- Questionnaire Data (including computed variables) for Sections A-K
    A. Background and demographic information
    B. Sex education, vasectomy & infertility, sexual intercourse, enumeration and relationship with up to 3 recent (or last) sexual partners
    C. Current wife or cohabiting partner: year of marriage; children; contraception with her
    D. Recent (or last) sexual partner(s) (up to three): key dates, children; contraception with her; 1st partner
    E. Former wives and first premarital cohabiting partner: year of key dates, children; contraception with each
    F. Other biological and adopted children; other pregnancies
    G. Fathering: Activities with the youngest child he lives with and the youngest child he lives apart from
    H. Desires and intentions for future children
    I. Health conditions, access to health care, and receipt of health services
    J. More background, more demographic information, & attitude questions
    K. Audio CASI: pregnancy reporting; cigarette, alcohol & other drug use; STD/HIV-risk behaviors; non-voluntary intercourse; sexual orientation & attraction; income and economic insecurity

Why I Chose this Dataset

- My team & I are trying to create an additional form of birth control for men
  - They currently only have condoms & vasectomies

THE THEORY ...

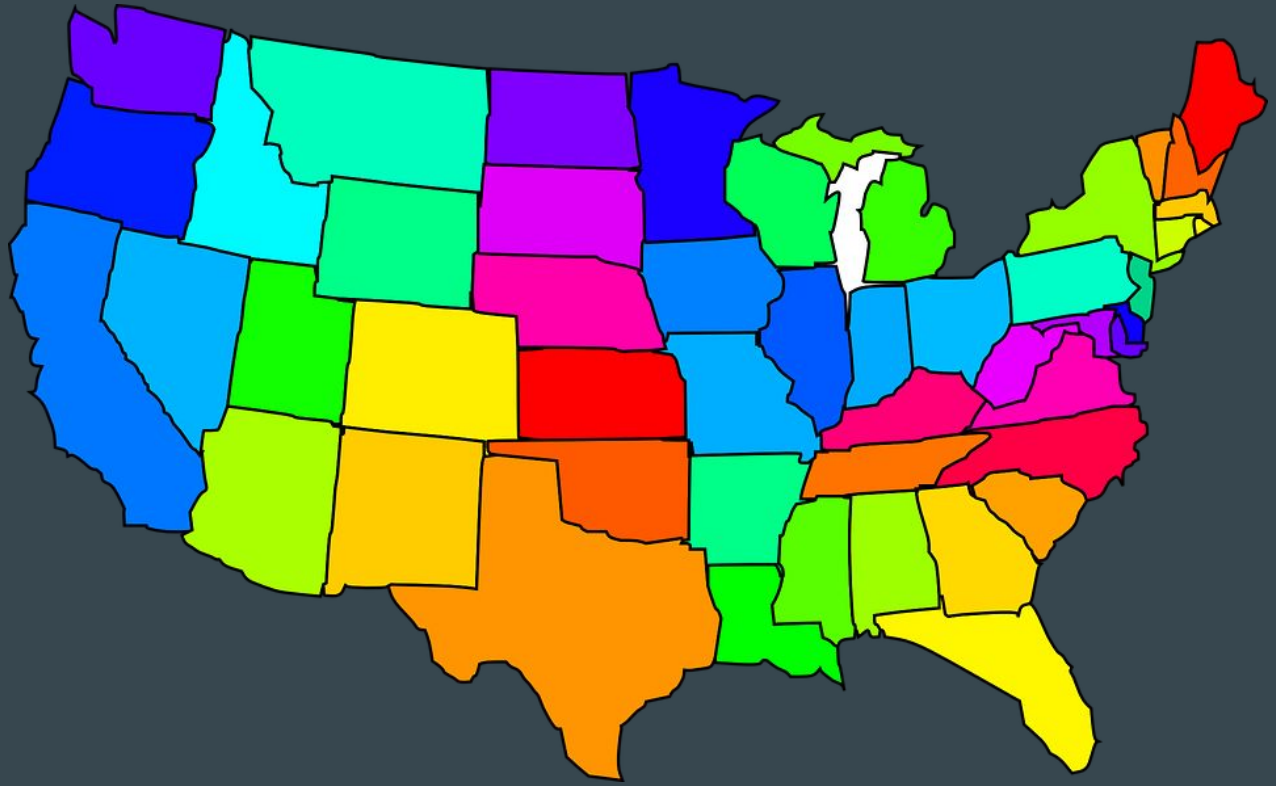
➔ Men who have children & have vasectomies are already apart of the family planning

➔ Men who play a role in family planning would be interested in a long-acting yet reversible form of contraception

What I am Trying to Learn

- How is the primary contraceptive for couples chosen?
- Where do men find information on birth control options?
- How is sex & procreation communicated between partners?
- Why do men get vasectomies?
  - After a certain number of children are they more likely to get one?
  - Are vasectomies more prominent around a certain age?
  - What are the demographics for these men who get a vasectomy?
- What men/ how likely are men to elect to get a vasectomy reversal procedure?

- Merging with Census Data
- Questioning whether vasectomies are more prevalent at certain geographical locations in the US



My Progress

- Written code for analysis:
  - **Basic stats** about my dataset
  - **Boxplot** showing distribution by age of men who get a vasectomy
  - **Boxplot** showing distribution by number of children a man has once he has decided to get a vasectomy
  - **Histogram** showing how many men rely on the different types of birth control
  - **Histogram** showing where men obtain their birth control information

- Still need debug code & ensure it runs properly

Analysis Of The Dataset

Below are codes and desciptions of the types of analysis methods I would run on my dataset. The goal is to try to understand why and when men elect to get vasectomies. Further anaylsis will be completed once I am able to load the dataset in an understandable way.

```
# Basic statistics about my dataset.
pd.DataFrame(2015_2017_MaleDat.vasectomy.describe())
```

```
# A boxplot showing the distribution by age of men who get a vasectomy.
f, ax = plt.subplots(figsize=(6.5, 6.5))
sns.boxplot(x="vesectomy", y="age", data= 2015_2017_MaleData, fliersize=0.5, linewidth=0.75, ax=ax)
```

```
# A histogram showing how many men rely on the different types of birthcontrol.
diamonds['birthcontrol'].hist(bins=np.arange(0,20000,2500))
plt.xlabel('Types of Birth Control')
plt.ylabel('Number of Men')
```

Merging Datasets

Because I could not properly load my main dataset, I can not actually merge my dataset. However, I found data that I would merge it with. I would use **data** from the US census to see if the men who elect to get vasectomies happen to be from a similar location. Therefore, my "foreign key" would be age (15-49 year old males). I would than plot the men of that age range who have had a vasectomy on spatial visualiztion map to analyze any trends.

```
import bq_helper
from bq_helper import BigQueryHelper
# https://www.kaggle.com/sohier/introduction-to-the-bq-helper-package
census_data = bq_helper.BigQueryHelper(active_project="bigquery-public-data",
                                       dataset_name="census_bureau_usa")
```

Deploying A Right Join Merge Of The Two Datasets

```
df1.merge(df2, left_on='2015_2017_MaleData', right_on='census_bureau_usa')
```

(I would than look for code examples to figure out the exact syntax I would need for my project to depict a map of the USA with specific areas highlighted where high number of vasectomy procedures are proformed.)

```
import geopandas as gpd
import pandas as pd
import pickle
import matplotlib.pyplot as plt
```

Encountered Problems

- Was not properly loading .dat file through pandas

	0	1	2	3	4	5	6	7	8	9	...	269	270
0	70626518141818181112	66	50	1	13512015	12	5	18155151	11	3	...	None	None
1	70629123532323235	1	36	50	1	18512011	12	122015	19155111	32	...	None	None
2	70631517531717175	1	46	50	1	115	11	16155111	43	5	...	None	None
3	70636537533737375	1	26	50	1	1412	12121996122003	1721151213141	5	135	...	None	None
4	70640549524949495	1	41	11115	12111985	12	113255111	31	3	5	...	None	None

- Found Stata and SAS files for the data
- Could not run them properly through pandas
- Used school computer Stata program to run files
- Stata file for loading the data was a “standard” file used for all datasets
    - Had to ensure the Male 2015-2017 files were properly called
- Found great help from the Data Lab at the Tisch Library!

- The school's computer lab version of Stata can only load datasets with number of variables < 2,040. I have 2,945 variables.

MALE RESPONDENT FILE *– Information for each male respondent*

- Respondent ID (CASEID) and selected screener variables
- Questionnaire Data (including computed variables) for Sections A-K
  - A. Background and demographic information
  - B. Sex education, vasectomy & infertility, sexual intercourse, enumeration and relationship with up to 3 recent (or last) sexual partners
  - C. Current wife or cohabiting partner: year of marriage; children; contraception with her
  - D. Recent (or last) sexual partner(s) (up to three): key dates, children; contraception with her; 1st partner
  - E. Former wives and first premarital cohabiting partner: year of key dates, children; contraception with each
  - F. Other biological and adopted children; other pregnancies ~~specifics~~
  - G. ~~Fathering: Activities with the youngest child he lives with and the youngest child he lives apart from~~
  - H. Desires and intentions for future children
  - I. Health conditions, access to health care, and receipt of health services
  - J. More background, more demographic information, & attitude questions
  - K. Audio CASI: pregnancy reporting; cigarette, ~~alcohol & other drug use; STD/HIV risk behaviors; non-voluntary intercourse;~~ sexual orientation & attraction; income and economic insecurity

Any Ideas

- What additional analysis could I try?
- What would you be interested in learning?

Group by? Correlation heat map?

Next Steps

1. Finish Uploading Data
2. Upload Dictionary
3. Run it Through Analysis Code
4. Finish Final Assignment :)